DOCUMENT RESUME

ED 392 817 TM 024 455

AUTHOR	Mislevy, Robert J.; And Others
TITLE	Toward a Test Theory for Assessing Student Understanding.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Naval Research, Arlington, VA. Cognitive and Neural Sciences Div.
REPORT NO	ETS-RR-91-32-ONR
PUB DATE	Apr 91
CONTRACT	N00014-88-K-0304
NOTE	46p.
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Cognitive Psychology; *Comprehension; *Constructivism (Learning); *Educational Assessment; Educational Practices; Educational Psychology; Item Response Theory; Knowledge Level; Learning Strategies; *Student Evaluation; Student Interests; *Test Theory
IDENTIFIERS	*Chunking

ABSTRACT

The view of learning that underlies standard test theory is inconsistent with the view rapidly emerging from cognitive and educational psychology. Learners become more competent not simply by learning more facts and skills, but by reconfiguring their knowledge; by "chunking" information to reduce memory loads; and by developing strategies and models that help them discern when and how facts and skills are important. Neither classical test theory nor item response theory is designed to inform educational decisions conceived from this perspective. This paper sketches the outlines of a test theory built around models of student understanding, as inspired by the substance and the psychology of the domain of interest. The ideas are illustrated with a simple numerical example based on the balance beam tasks of R. S. Siegler. Directions in which the approach must be developed to be broadly useful in educational practice are discussed. An appendix gives equations for the latent class model. (Contains 2 tables, 8 figures, and 48 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 392 817

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

RR-91-32-ONR

TOWARD A TEST THEORY FOR ASSESSING STUDENT UNDERSTANDING

Robert J. Mislevy
Kentaro Yamamoto
and
Steven Anacker

This research was sponsored in part by the
Program Research Planning Council
Educational Testing Service; and the
Cognitive Science Program
Cognitive and Neural Sciences Division
Office of Naval Research, under
Contract No. N00014-88-K-0304
R&T 4421552

Robert J. Mislevy, Principal Investigator

Educational Testing Service
Princeton, New Jersey

April 1991

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

Approved for public release; distribution unlimited.

TM024455

BEST COPY AVAILABLE

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION: Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): RR-91-32-ONR	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION: Educational Testing Service; 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION: Cognitive Science Program, Office of Naval Research (Code 1142CS), 800 North Quincy Street
6c ADDRESS (City, State, and ZIP Code): Princeton, NJ 08541	7b ADDRESS (City, State, and ZIP Code): Arlington, VA 22217-5000
8a NAME OF FUNDING/SPONSORING ORGANIZATION; 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-88-K-0304
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO: 61153N; PROJECT NO: RR04204; TASK NO: RR04204-01; WORK UNIT ACCESSION NO: R&T4421552

11 TITLE (Include Security Classification)

Toward a Test Theory for Assessing Student Understanding (Unclassified)

12 PERSONAL AUTHOR(S)

Robert J. Mislevy, Kentaro Yamamoto, and Steven Anacker

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Technical	FROM _____ TO _____	April 1991	37

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
05	10		Inference networks, cognitive assessment, student models, Bayesian inference

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The view of learning that underlies standard test theory is inconsistent with the view rapidly emerging from cognitive and educational psychology. Learners become more competent not simply by learning more facts and skills, but by reconfiguring their knowledge; by "chunking" information to reduce memory loads; and by developing strategies and models that help them discern when and how facts and skills are important. Neither classical test theory nor item response theory (IRT) is designed to

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL: Dr. Charles E. Davis	22b TELEPHONE (Include Area Code): 703-696-4046; 22c OFFICE SYMBOL: ONR 1142CS

DD Form 1473, JUN 86 — Previous editions are obsolete — S/N 0102-LF-014-6603

BEST COPY AVAILABLE

19 ABSTRACT

inform educational decisions conceived from this perspective. This paper sketches the outlines of a test theory built around models of student understanding, as inspired by the substance and the psychology of the domain of interest. The ideas are illustrated with a simple numerical example based on Siegler's balance beam tasks. Directions in which the approach must be developed to be broadly useful in educational practice are discussed.

Toward a Test Theory for Assessing Student Understanding

Robert J. Mislevy, Kentaro Yamamoto, and Steven Anacker

Educational Testing Service

April, 1991

This work was supported by Contract No. N00014-88-K-0304, R&T 4421552, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research, and by the Program Research Planning Council of Educational Testing Service. We are grateful to Dr. Steen Andreassen for permission to reproduce Figures 1 and 2, to Robert Siegler for the use of data from his balance-beam studies, and to Larry Frase and Kikumi Tatsuoka for comments on an earlier version of the paper.

Toward a Test Theory for Assessing Student Understanding

Robert J. Mislevy, Kentaro Yamamoto, and Steven Anacker

Educational Testing Service

Abstract

The view of learning that underlies standard test theory is inconsistent with the view rapidly emerging from cognitive and educational psychology. Learners become more competent not simply by learning more facts and skills, but by reconfiguring their knowledge; by "chunking" information to reduce memory loads; and by developing strategies and models that help them discern when and how facts and skills are important. Neither classical test theory nor item response theory (IRT) is designed to inform educational decisions conceived from this perspective. This paper sketches the outlines of a test theory built around models of student understanding, as inspired by the substance and the psychology of the domain of interest. The ideas are illustrated with a simple numerical example based on Siegler's balance beam tasks. Directions in which the approach must be developed to be broadly useful in educational practice are discussed.

Background

When schooling became mandatory at the turn of the century, educators suddenly faced selection and placement decisions for unprecedented numbers of students, displaying the diversity of abilities and backgrounds that individuals bring to schooling (Glaser, 1981). Numbers of correct answers to multiple-choice test items were used to rank students according to their overall proficiencies in domains of tasks. These rankings were used in turn to predict students' success in fixed educational experiences.

Classical test theory (CTT) emerged when Spearman (e.g., 1907) applied statistical methods to study how reliable estimates of this overall proficiency would be from different test forms that might be constructed for the purpose. Extensions of this work led over the years to a vast armamentarium of techniques for building tests and making decisions with test scores (Gulliksen, 1950); to an axiomatic foundation for statistical inference about test scores (Lord, 1959; Lord & Novick, 1968; Novick, 1966); and to sophisticated techniques for partitioning test score variance according to facets of items, persons, and observational settings (Cronbach, Gleser, Nanda, & Rajaratnam,1972). It is important to note that in all this work, the object of inference is overall proficiency—the test score, observed or expected— in terms of numbers of correct responses in a domain of items.

Item response theory (IRT; see Hambleton, 1989, for an overview) represented a major practical advance over CTT by modeling probabilities of correct item response in terms of an unobservable proficiency variable. IRT solves many problems that were difficult under CTT, in equating, test construction, and adaptive testing. Advanced statistical methods have been brought to bear on inferential problems in IRT, including sophisticated estimation algorithms (e.g., Bock & Aitkin, 1981), techniques from missing-data theory (Mislevy, in press-a), and Bayesian treatments of uncertainty in models and parameters (Lewis, 1985; Mislevy & Sheehan, 1990; Tsutakawa & Johnson, 1988). The underlying psychological model remains quite simple, however; as in CTT, the focus remains on overall proficiency in a domain of items. From the perspective of IRT, two students with the same overall proficiency are indistinguishable.

As useful as standard tests and standard test theory have proven in large-scale evaluation, selection, and placement problems, their focus on *who* is competent and *how many* items they answer can fall short when the goal is to improve individuals' competencies. Glaser, Lesgold, and Lajoie (1987) point out that tests can predict failure

without an understanding of what causes success, but intervening to prevent failure and enhance competence requires deeper understanding.

The past decade has witnessed considerable progress toward the requisite understanding. Psychological research has moved away from the traditional laboratory studies of simple (even random!) tasks, to tasks that better approximate the meaningful learning and problem-solving activities that engage people in real life. Studies comparing the ways experts differ from novices in applied problem-solving in domains such as physics and trouble-shooting (e.g., Chi, Feltovich & Glaser, 1981) reveal the central importance of knowledge structures—networks of concepts and interconnections among them—that impart meaning to patterns in what one observes and how one chooses to act. The process of learning is to a large degree expanding these structures and, importantly, *reconfiguring them* to incorporate new and qualitatively different connections as the level of understanding deepens. Educational psychologists have begun to put these findings to work in designing both instruction and tests (e.g., Glaser et al., 1987; Greeno, 1976; Marshall, 1985, in press). Again in the words of Glaser, Lesgold, and Lajoie (1987),

> "Achievement testing as we have defined it is a method of indexing stages of competence through indicators of the level of development of knowledge, skill, and cognitive process. These indicators display stages of performance that have been attained and on which further learning can proceed. They also show forms of error and misconceptions in knowledge that result in inefficient and incomplete knowledge and skill, and that need instructional attention." (p.81)

Paraphrasing Ohlsson and Langley (1985), Clancey (1986) summarizes the shift in perspective: "[to] describing mental processes, rather than quantifying performance with respect to stimulus variables; describing individuals in detail, not just stating generalities; and giving psychological interpretation to qualitative data, rather than statistical treatment to numerical measurements" (p. 391).

An Approach to Modeling Student Understanding

The modeling approach we are beginning to pursue can be encapsulated as follows:

> "Standard test theory evolved as the application of statistical theory with a simple model of ability that suited the decision-making environment of mass educational systems. Broader educational options, based on insights into

> the nature of learning and supported by more powerful technologies, demand a broader range of models of capabilities—still simple compared to the realities of cognition, but capturing patterns that inform a broader range of instructional alternatives. A new test theory can be brought about by applying to well-chosen cognitive models the same general principles of statistical inference that led to standard test theory when applied to the simple model." (Mislevy, in press-b).

The approach begins in a specific application by defining a universe of student models. This "supermodel" is indexed by parameters that signify distinctions between states of understanding. Symbolically, we shall refer to the (typically vector-valued) parameter of the student-model as η. A particular set of values of η specifies a particular student model, or one particular state among the universe of possible states the supermodel can accommodate. These parameters can be qualitative or quantitative, and qualitative parameters can be unordered, partially ordered, or completely ordered. A supermodel can contain any mixture of these types. Their nature is derived from the structure and the psychology of the learning area, the idea being to capture the essential distinctions among students.

Any application faces a modeling problem, an item construction problem, and an inference problem.

The *modeling* problem is delineating the states or levels of understanding in a learning domain. In meaningful applications this might address several distinct strands of learning, as understanding develops in a number of key concepts, and it might address the connectivity among those concepts.[1] Symbolically, this substep defines the *structure* of $p(\mathbf{x}|\eta)$, where $\mathbf{x}$ represents observations. Obviously any model will be a gross simplification of the reality of cognition. A first consideration in what to include in the supermodel is the substance and the psychology of the domain: Just what are the key

---

[1] A particularly interesting special case occurs when the universe of student models can be expressed a performance models (Clancey, 1986). A performance model consists a knowledge base and manipulation rules that can be run on problems in a domain of interest. A particular model can contain both knowledge and production rules that are incorrect or incomplete; the solutions it produces will be correct or incorrect in identifiable ways. Here the parameter η specifies features of performance models.

concepts? What are important ways of understanding and misunderstanding them? What are typical paths to competence? A second consideration is the so-called grain-size problem, or the level of detail at which student-models should differ. A major factor in answering this question is the decision-making framework under which the modeling will take place. As Greeno (1976) points out, "It may not be critical to distinguish between models differing in processing details if the details lack important implications for quality of student performance in instructional situations, or the ability of students to progress to further stages of knowledge and understanding."

The *item construction* problem is devising situations for which students who differ in the parameter space are likely to behave in observably different ways. The conditional probabilities of behavior of different types given the unobservable state of the student are the *values* of $p(\mathbf{x}|\eta)$, which may in turn be modeled in terms of another set of parameters, say β. The $p(\mathbf{x}|\eta)$ values provide the basis for inferring back about the student state. An element in $\mathbf{x}$ could contain a right or wrong answer to a multiple-choice test item, but it could instead be the problem-solving approach regardless of whether the answer is right or wrong, the quickness of a responding, a characteristic of a think-aloud protocol, or an expert's evaluation of a particular aspect of the performance. The effectiveness of an item is reflected in differences in conditional probabilities associated with different parameter configurations, so an item may be very useful in distinguishing among some aspects of potential student models but useless for distinguishing among others. Tatsuoka (1989) demonstrates the relationship between item construction and inference about students' strategies for subtracting mixed numbers.

The *inference* problem is reasoning from observations to student models. The model-building and item construction steps provide η and $p(\mathbf{x}|\eta)$. Let $p(\eta)$ represent expectations about η in a population of interest—possibly non-informative, possibly based on expert opinion or previous analyses. Bayes theorem can be employed to draw inferences about η given $\mathbf{x}$ via $p(\eta|\mathbf{x}) \propto p(\mathbf{x}|\eta)\, p(\eta)$. Thus $p(\eta|\mathbf{x})$ characterizes belief about a particular student's model after having observed a sample of the student's behavior. Practical problems include characterizing what is known about β so as to determine $p(\mathbf{x}|\eta)$, carrying out the computations involved in determining $p(\eta|\mathbf{x})$, and, in some applications, developing strategies for efficient sequential gathering of observations. As we have noted, analogous problems have been studied in standard test theory, and the solutions there, because they are applications of general principles of statistical inference, generalize to

models built around alternative psychological models. The models are more realistic and more ambitious, but the formalism is identical.[2]

Previous Research

Research relevant to this approach has been carried out in a wide variety of fields, including cognitive psychology, the psychology of mathematics and science education, artificial intelligence (AI) work on student modeling, test theory, and statistical inference. Cognitive scientists have suggested general structures such as "frames" or "schemas" that can serve as a basis for modeling understanding (e.g., Minsky, 1975; Rumelhart, 1980), and have begun to devise tasks that probe their features (e.g., Marshall, 1989, in press). Researchers interested in the psychology of learning in subject areas such as proportional reasoning have focused on identifying key concepts, studying how they are typically acquired (e.g., in mechanics, Clement, 1982; in ratio and proportional reasoning, Karplus, Pulos, & Stage, 1983), and constructing observational settings that allow one to infer students' understanding (e.g., van den Heuvel, 1990; McDermott, 1984). We make no effort here to review these literatures, but point out that our work can succeed only by building upon their foundations. Our potential contribution would be to the structures and mechanics of model-building and inference. The following sections briefly mention some important work along these lines from test theory and statistics.

Modeling Student Behavior

The standard models of educational measurement are concerned solely with examinees' tendencies to answer items correctly—that is, their overall proficiency. Recently, however, models that focus on patterns other than overall proficiency have begun to appear the test theory literature. Some examples that are relevant to educational applications are listed below.

---

[2] Advocates of student modeling emphasize the qualitative aspects of student models. Our approach is compatible with this view, as it is possible to build universes of qualitative models, indexed by parameters that distinguish their features. Our knowledge about a particular student's model is imperfect, however. It can be expressed in terms of probabilities expressing the plausibility of various models, given what has been observed. Probabilities are quantitative, and admit to a calculus of manipulation. We might thus employ a *quantitative* model for our (imperfect) knowledge about *qualitative* student models.

1. Mislevy and Verhelst's (1990) *mixture models* for item responses when different examinees follow different solution strategies or use alternative mental models. When a single IRT model cannot capture key distinctions among examinees, it may suffice to posit qualitatively distinct classes of examinees and use IRT models to summarize distinctions among examinees within these classes.

2. Wilson's (1989b) *Saltus* model for characterizing stages of conceptual development. This model parameterizes the differential patterns of strength and weakness expected as learners progress through successive conceptualizations of a domain.

3. Falmagne's (1989) and Haertel's (1984) latent class models for *Binary Skills*. These models are intended for domains in which competence can be described by the presence or absence of several (possibly complex) elements of skill or knowledge, and observational situations can be devised that demand various combinations of these skills. Also see Paulson (1986) for an alternative use of latent class modelling in cognitive assessment.

4. Embretson's (1985) *multicomponent models* for integrating item construction and inference within a unified cognitive model. The conditional probabilities of solution steps given a multifaceted student model are given by IRT-like statistical structures.

5. Tatsuoka's (1989) *Rule space* analysis. Tatsuoka uses a generalization of IRT methodology to define a metric for classifying examinees based on likely patterns of item response given patterns of knowledge and strategies.

6. Yamamoto's (1987) *Hybrid* model for dichotomous responses. The *Hybrid* model characterizes an examinee as either belonging to one of a number of classes associated with states of understanding, or in a catch-all IRT class. This approach might be useful when certain response patterns signal states of understanding for which particular educational experiences are known to be effective. Instructional decisions are triggered by these patterns if they are detected, but by overall proficiency when no more targeted action can be provided.

7. Masters and Mislevy's (in press) and Wilson's (1989a) use of the *Partial Credit* rating scale model to characterize levels of understanding, as evidenced by the nature or approach of a performance rather than its correctness. These applications incorporate into a

probabilistic framework the cognitive perspective underlying Biggs and Collis's (1982) SOLO taxonomy for describing salient qualities of performances.

These are the rudiments of models upon which concept-referenced achievement measures can be based. Applications to date have been fairly limited, and most have addressed one-to-many relationships between an underlying knowledge state and observable behavior. That is, a single (possibly unordered or multifaceted) variable has been used to characterize examinees, and performance on all items is modeled in terms of this variable. What is lacking from the point of view of the educator is the fact that meaningful real world tasks are rarely segregated into these neat little sets. Rather, they often involve multiple concepts, connections among larger concepts, and transformations among alternative representations of a domain. While the simple tasks that characterize one-to-many domains are essential at early stages of learning, more complex tasks that involve multiple concepts in many-to-many relationships are needed to promote the integration among concepts that form the core of what is often called "higher-level learning."

Inference Networks

Recent developments in the context of probability-based inference networks (Lauritzen & Spiegelhalter, 1988; Pearl, 1988) offer a capability for integrating conceptual models of the type described above. These probability-based structures are attractive for educational measurement because they permit a coherent extension of the modeling approach and inferential logic of the new cognitive-assessment models mentioned above. To show how the approach might be applied in the educational setting, we first discuss an application in the setting of medical diagnosis.

MUNIN is an inference network that organizes knowledge in the domain of electromyography—the relationships among nerves and muscles. Its function is to diagnose nerve/muscle disease states. The interested reader is referred to Andreassen, Woldbye, Falck, and Andersen (1987) for a fuller description. The prototype discussed in that presentation and used for our illustration concerns a single arm muscle, with concepts

represented by twenty-five nodes and their interactions represented by causal links.[3] A graphic representation of the network appears in Figure 1.

[Figure 1 about here]

The rightmost column of nodes in Figure 1 concerns outcomes of potentially observable variables, such as symptoms or test results. These outcomes are the **x** vector in our earlier notation. The middle layers are "pathophysiological states," or syndromes. These drive the probabilities of observations. The leftmost layer is the underlying disease state, including three possible diseases in various stages, no disease, or "Other"—a condition not built into in the system. These states drive the probabilities of syndromes. It is assumed that a patient's true state can be adequately characterized by values of these disease and syndrome states—our η parameter. Paths indicate conditional probability relationships, which are to be determined either logically, subjectively, purely empirically, or through model-based statistical estimation. In particular, the paths ending at observables represent $p(\mathbf{x}|\eta)$. Note that the probabilities of observables depend on some syndromes, but not others. The lack of a path signifies conditional independence. Note also that a given test result can be caused by different disease combinations.

As a patient enters the clinic, the diagnostician's state of knowledge about him is expressed by population base rates, or $p(\eta)$. This is depicted in Figure 1 by bars that represent the base probabilities of disease and syndrome states. Base rates of observable test results are similarly shown. Tests are carried out, one at a time or in clusters, and with each result the probabilities of disease states are updated. The expectations of tests not yet given are calculated, and it can be determined which test will be most informative in identifying the disease state. Knowledge is thus accumulated in stages, from $p(\eta)$ to $p(\eta|\mathbf{x}_1)$ after observing the first subset of tests, to $p(\eta|\mathbf{x}_1,\mathbf{x}_2)$ after the second, and so on, with each successive test selected optimally in light of knowledge at that point in time. Figure 2 illustrates the state of knowledge after a number of electromyographic test results have been observed. Observable nodes with results now known are depicted with shaded bars representing observed values For them, knowledge is perfect. The implications of these results have been propagated leftward to syndromes and disease states, as shown by

---

[3] The ESPRIT team has generalized the application to address clusters of interrelated muscles in a network containing over a thousand nodes.

distributions that differ from the base rates in Figure 1. These values guide the decision to test further or initiate a treatment. Finally, updated beliefs about disease states have been propagated back toward the right to update expectations about the likely outcomes of test not yet administered. These expectations, and the potential they hold for further updating knowledge about the disease states, guide the selection of further tests.

[Figure 2 about here]

Inference Networks in the Educational Setting

To see how the ideas underlying MUNIN apply to the educational setting, consider the following analogy:

Medical Application	Educational Application
Observable symptoms, medical tests	Test items, verbal protocols, teachers' ratings of levels of understanding, solution traces
Disease states, syndromes	States or levels of understanding of key concepts, available strategies
Architecture of interconnections based on medical theory	Architecture of interconnections based on cognitive and educational theory
Conditional probabilities given by physiological models, empirical data, expert opinion	Conditional probabilities given by psychological models, empirical data, expert opinion

The definitions of key concepts will be guided by theorized and observed stages of learning in the area, and the connections with observables will be expressed through measurement models such as those discussed above. The initialization of the probabilities in the network will be accomplished by one or more methods: clinical analysis, with skilled interviewers assessing in detail the nature of students' understandings and related these understandings to task performances; statistical analysis of data concerning selected models

for portions of the larger network (Mislevy & Verhelst, 1990); or theoretical analysis, in which logic or theory provides expectations for outcomes under hypothesized cognitive states. After the initialization phase, connections can be updated periodically with the larger amounts of less precise data that will be accumulated as students provide information about the adequacy of the relationships embodied in the network and the accuracy of the baseline and conditional probabilities.

A Numerical Example

Siegler's balance beam tasks

Kuhn (1970) emphasizes the central role that exemplars, or small, archetypical examples, play in science. Textbook examples are the vehicle through which students are acculturated to the concepts and relationships of a particular way of viewing a class of phenomena—a paradigm, in Kuhn's words. They function almost like parables or morality tales. New paradigms are introduced with new exemplars, that introduce new concepts, highlight differences between the new paradigm and the old, and demonstrate how the new way of thinking solves problems the old way could not. Modeling the states of the electron in the hydrogen atom possesses this status in quantum mechanics. Explaining children's understanding of balance beam problems, an exemplar from developmental psychology originated by Piaget, is approaching the same status in test theory (e.g., Kempf, 1983, Mislevy, in press-b, and Wilson, 1989b). Robert Siegler's balance beam tasks yield data that are, on the surface, indistinguishable from standard test data, but there are two key distinctions:

1. What is important about examinees is not their overall probability of answering items correctly, but their (unobservable) state of understanding of the domain.

2. Children at less sophisticated levels of understanding initially get certain problems right for the wrong reasons. These items are more likely to be answered wrong at intermediate stages, as understanding deepens! They are bad items by the standards of classical test theory and IRT, because probabilities of correct response do not increase monotonically with increasing total test score. From the perspective of the developmental theory, however, not only is this reversal expected, but it plays an important role in distinguishing among children with different ways of thinking about the problems.

Attempting to study children's reasoning in a manner less subjective than Piaget's unstructured interviews, Siegler (1981) devised a series of balance beam tasks like the one illustrated in Figure 3. Varying numbers of weights are placed at varying locations on a balance beam. The child predicts whether the beam will tip to left, to the right, or remain in balance. Piaget's analysis of children's behavior on balancing tasks (Inhelder & Piaget, 1958), posits that a child will respond in accordance with his or her stage of understanding. The usual stages through which children progress can be described in terms of successive acquisition of the rules listed below.

[Figure 3 about here]

Rule I: If the weights on both sides are equal, it will balance. If they are not equal, the side with the heavier weight will go down. (Weight is the "dominant dimension," because children are generally aware that weight is important in the problem earlier than they realize that distance from the fulcrum, the "subordinate dimension," also matters.)

Rule II: If the weights and distances on both sides are equal, then the beam will balance. If the weights are equal but the distances are not, the side with the longer distance will go down. Otherwise, the side with the heavier weight will go down. (A child using this rule uses the subordinate dimension only when information from the dominant dimension is equivocal.)

Rule III: Same as Rule II, except that if the values of both weight and length are unequal on both sides, the child will "muddle through" (Siegler, 1981, p.6). (A child using this rule now knows that both dimensions matter, but doesn't know just how they combine. Responses will be based on a strategy such as guessing.)

Rule IV: Combine weights and lengths correctly (i.e., compare torques, or products of weights and distances).

It was thus hypothesized that each child could be classified into one of five stages—the four characterized by the rules, or an earlier "preoperational" stage in which neither weight nor length are thought to bear any systematic relationship to the action of the beam.

Siegler developed six types of problems listed below to distinguish among children at different stages of reasoning. (See Figure 4 for an example of each.)

Equal problems (E), with matching weights and lengths on both sides.

Dominant problems (D), with unequal weights but equal lengths.

Subordinate problems (S), with unequal lengths but equal weights.

Conflict-dominant problems (CD), in which one side has greater weight, the other has greater length, and the side with the heavier weight will go down.

Conflict-subordinate problems (CS), in which one side has greater weight, the other has greater length, and the side with the greater length will go down.

Conflict-equal problems (CE), in which one side has greater weight, the other has greater length, and the beam will balance.

[Figure 4 about here]

Table 1 shows the probabilities of correct response that would be expected from groups of children in different stages, if their responses were in complete accordance the hypothesized rules. Scanning across the rows reveals how the probability of a correct response to a given type of item does not always increase as level of understanding increases. For example, Stage II children tend to answer CD items right for the wrong reason, while Stage III children, now aware of a conflict, flounder.

[Table 1 about here]

A latent class model for balance beam tasks

If the theory were perfect, the columns in Table 1 would give probabilities of correct response to the various types of items from children at different stages of understanding. Observing a correct response to an S item, for example, would eliminate the possibility that the child was in Stage I. But because the model is not perfect[4], and because children make slips and lucky guesses, any response could be observed from a child in any stage. A latent class model (Lazarsfeld, 1950) can be used to express the

---

[4] This model assumes that the five states are exhaustive and mutually exclusive. Alternative models, such as those of Tatsuoka and Yamamoto mentioned earlier, could be used to relax these restrictions.

structure posited in Table 1 while allowing for some "noise" in real data (see Appendix for details). Instead of expecting incorrect responses with probability one to S items from Stage I children, we might posits some small fraction of correct answers—p(S correct| Stage=I). Similar probabilities of "false positives" can be estimated for other cells in Table 1 containing 0's. In the same spirit, probabilities less than one, due to "false negatives," can be estimated for the cells with 1's. Note that inferences cannot be as strong when these uncertainties are present; a correct response to an S item still suggests that a child is probably not in Stage I, but no longer is it proof positive.

Expressing this model in the notation introduced above, η represents stage membership, $\mathbf{x}$ represents item responses, and $p(\mathbf{x}|\eta)$ are conditional probabilities of correct responses to items of the various types from children in different stages—a noisy version of Table 1. The proportions of children in a population of interest at the different stages are $p(\eta)$, and the probabilities that convey our knowledge about a child's stage after we have observed his responses are $p(\eta|\mathbf{x})$.

Siegler created a 24-task test comprised of four tasks of each type. He collected data from 60 children, from age 3 up through college age, at two points in time, for a total of 120 response vectors. We fit a latent class model to these data using the HYBRIL computer program (Yamamoto, 1987), obtaining the conditional probabilities—$p(\mathbf{x}|\eta)$—shown in Table 2, and the following vector summarizing the (estimated) population distribution of stage membership:

$$p(\eta) = (\text{Prob(Stage=0)}, \text{Prob(Stage=I)}, ..., \text{Prob(Stage=IV)})$$

$$= (.257,.227,.163,.275,.078) .$$

[Table 2 about here]

Note that different types of items are differentially useful to distinguish among children at different levels. E items, for example, are best for distinguishing Stage 0 children from everyone else. CD items, which would be dropped from standard tests because their probabilities of correct response do not have a strictly increasing relationship with total scores, help differentiate among children at Stages II, III, and IV.

Figure 5 depicts the state of knowledge about a child before observing any responses using the conventions of the MUNIN figures. Just one item of each type is shown rather than all four for simplicity. The corresponding status of an observable node

(i.e., an item type) is the expectation of a correct response from a child selected at random from the population. The path from the stage-membership node to a particular observable node represents a row of Table 2.

[Figure 5 about here]

Adaptive testing

Figure 5 represents the state of our knowledge about a child's reasoning stage and expected responses before any actual responses are observed. How does knowledge change when a response is observed? One of the children in the sample, Douglas, gave an incorrect response to his first S item. This could happen regardless of Douglas' true stage; the probabilities are obtained by subtracting the entries in the S row of Table 2 from 1.000, yielding, for Stages 0 through IV, .667, .973, .116, .019, and .057 respectively. This is the likelihood function for η induced by the observation of the response. The bulk of the evidence is for Stages 0 and I. Combining these values with the initial stage probabilities $p(\eta)$ via Bayes theorem yields updated stage probabilities, $p(\eta|$incorrect response to an S item): for Stages 0 through IV respectively, .41, .52, .04, .01, and .01. Expectations for items not yet administered also change. They are averages of the probabilities of correct response expected from the various stages, now weighted by the new stage membership probabilities. The state of knowledge after observing Douglas' first response is depicted in Figure 6 (see Appendix for details; also see Macready & Dayton, 1989.)

[Figure 6 about here]

In a simulation of adaptive testing, we updated our knowledge about Douglas one response at a time, at each step looking at his actual response to an item expected to most substantially reduce our uncertainty about his stage membership. Figure 7 charts probabilities of stage membership for Douglas after each of the first ten items, showing that we quickly converge to Stage 0.

[Figure 7 about here]

Extending the paradigm

The balance beam exemplar illustrates the challenge of inferring states of understanding, but it addresses development of only a single key concept. A major thrust of our proposal is to characterize interconnections among distinct lines of development.

This section takes a small step in this direction by discussing a hypothetical extension to the exemplar, namely, the ability to carry out the arithmetic operations needed to calculate torques. For illustrative purposes, we simply posit a skill to carry these calculations out reliably, either possessed by a child or not. Obviously states of understanding could be developed in greater detail here.

Calculating and comparing torques to solve the "conflict" problems characterizes Stage IV. But if a child at Stage IV cannot carry out the calculations reliably, his pattern of correct and incorrect responses would be hard to distinguish from that of a child in Stage III. Although the two children might answer about the same number of items correctly, the instruction appropriate for them would differ dramatically. And children at any stage of understanding of the balance beam might be able to carry out the computational operations in isolation. The goal of the extended system is to infer both balance-beam understanding and computational skill. To make the distinctions among states of understanding in this extended domain, we introduce two new types of observations:

1. Items isolating computation, such as "Which is greater, 3x4 or 5x2?"

2. Probes for introspection about solutions to conflict items: "How did you get your answer?"

Figure 8 offers one possible structure for this network. Others could be entertained, and in practice one would compare the degree to which they accord with observed data. To keep the diagram simple, only one balance-beam task each for an S and a CS task are illustrated. E and D items would have the same paths as the S task, and CD and CE tasks would have the same paths as the CS tasks. Also, the paths from Stage 0, I, and II indicators to balance beam tasks are not drawn in. The *structure* of paths, but not necessarily the *values*, would be the same as those connecting the Stage III indicator to those tasks.

[Figure 8 about here]

There are three kinds of unobservable variables in the system. The first group expresses level of understanding in the balance beam domain. It proves convenient to express stage membership in terms of dichotomous indicator variables for each stage, because of the special relationship of Stage IV to computational skill. Second is the ability to carry out the calculations involved in computing torques. The third concerns the integration of balance-beam understanding and calculating proficiency. Specifically, we

posit an indicator for whether a child both is in Stage IV *and* possesses the requisite computational skills. Other features of the network worth mentioning are as follows.

1. The probabilities of the pure computation items depend on the unobservable computation variable only; they are conditionally independent of level of balance beam understanding.

2. The correctness aspect of an answer has only two possibilities, right or wrong, but an explanation can fall into five categories corresponding to levels of understanding. A Stage III child might give an explanation consistent with Stages 0, I, II, or III, but would not give a Stage IV explanation. Theory thus posits that the conditional probability of a Stage K response from a Stage J child is zero if K>J. Conditional probabilities for K≤J might be estimated from data or based on experts' experience. It may turn out, for example, that the most likely explanation for an E task from people at Stage IV would probably be a Stage II explanation: "It balances because both the weights and distances are equal."

3. For children in Stages 0 through III, both the right/wrong answers and the "How" answers to balance beam tasks depend only on level of understanding. Because they do not realize the connection between the problems and the torque calculations, their responses to the balance beam tasks are conditionally independent of their computational skill, even on items for which that skill is an integral component of an expert solution.

4. For children in Stage IV, right/wrong answers to conflict items depend on the understanding/computation integration variable, but "How" answers depend only on understanding. A child in Stage IV with low computational skill can thus be differentiated from a child in Stage III by his higher probabilities of giving Stage IV explanations and incorrect answers to pure computation problems.

Discussion

This conceptual framework described above holds the promise of extending and clarifying standard educational measurement practices in several ways:

Connections with instruction can be forged more easily than with standard tests, because the focus is no longer on *how many* questions a student can answer, but *how* they answer them. In medical diagnosis, different diseases gave rise to similar results in certain

tests; in education, so too can different approaches lead to similar test scores for students. But accounting for the patterns of performance, especially if probing adaptively, can pinpoint the areas which need attention to best improve performance.

Student reports can be provided at varying levels and highlighting different features of a student's status. Of particular importance to the student and the teacher are reports in terms of levels or stages of understanding of key concepts, since this is the level at which instruction is aimed. For the quality control purposes of administrators, however, one could predict a student's performance on a standard set of tasks in the domain—say, a "market basket" of tasks that, ideally, every student should eventually be able to handle.

Use of different strategies or mental models can be accommodated in an inference network. This can take the form of either a single strategy/mental model choice for all tasks in a class, as studied by Mislevy and Verhelst (1990), or strategy/model switching from one task to another (as in Snow & Lohman, 1984). The nature and the strength of inferences one can draw will depend on the potential observational settings. With rich information, such as verbal protocols or partial solutions, it may be possible to characterize the range of solution methods the student has available and the conditions under which he employs them.

Testing "higher-order thinking" can be accomplished by including unobservable nodes for connections among more basic facts or concepts, and observable nodes that correspond to tasks for which the relationships of interest are critical. Because such tasks might well be open-ended and approachable in a variety of ways, the possibility of alternative solution strategies would need to be built into the network.

Adaptive testing can be carried out among concepts, not just for a single concept. IRT applications of adaptive testing are based on the one-to-many relationships that are appropriate for determining overall levels of proficiency, but inadequate for understanding connections among concepts. The inference network facilitates stepping variously throughout a domain, gathering information about critical domains by presenting tasks that call for varying combinations of key skills.

Handling atypical knowledge configurations or observational patterns can be accomplished by incorporating nodes analogous to the "Other" disease state in MUNIN or the catch-all IRT class in Yamamoto's (1987) *Hybrid* model. An "Other" state of understanding is a mechanism for capturing observational patterns that do not accord with

those specifically built into the network. A situation-sensitive student report might be generated in an instructional system when such a node becomes prominent, signalling that more intelligence than is embodied in the system is needed to figure out what this student is doing, and decide what to do about it.

Conclusion

Learning can be enhanced by a unified conceptual framework for instruction, testing, and reporting, because only in such a framework can coherent feedback loops be constructed. This presentation has focused on the educational measurement aspect of a system built on this premise. The recent introduction of measurement models built around states of understanding, and of inferential techniques to connect such pieces into networks that describe domains of school learning, provide a foundation for improved educational practice in this manner.

References

Andreassen, S., Woldbye, M., Falck, & Andersen, S.K. (1987). MUNIN: A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence.*

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Clancey, W.J. (1986). Qualitative student models. *Annual Review of Computer Science, 1*, 381-450.

Clement, J. (1982). Students' preconceptions of introductory mechanics. *American Journal of Physics, 50*, 66-71.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Falmagne, J-C. (1989). A latent trait model via a stochastic learning theory for a knowledge space. *Psychometrika, 54*, 283-303.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*, 923-936.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing* (Vol. 3). Hillsdale, NJ: Erlbaum.

Greeno, J.G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), *Cognition and instruction*. Hillsdale, NJ: Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Haertel, E.H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8*, 333-346.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.). New York: American council on Education/Macmillan.

van den Heuvel, M. (1990). Realistic arithmetic/mathematics instruction and tests. In K. Gravemeijer, M. van den Heuvel, & L. Streefland (Eds.), *Context free productions tests and geometry in realistic mathematics education*. Utrecht, The Netherlands: Research Group for Mathematical Education and Educational computer Center, State University of Utrecht.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic.

Karplus, R., Pulos, S., & Stage, E. (1983). Proportional reasoning of early adolescents. In R.A. Lesh & M. Landau (Eds.), Acquisition of mathematics concepts and processes. Orlando, FL: Academic Press.

Kempf, W. (1983). Some theoretical concerns about applying latent trait models in educational testing. In S.B. Anderson & J.S. Helmick (Eds.), *On educational testing*. San Francisco: Josey-Bass.

Kuhn, T.S. (1970). *The structure of scientific revolutions* (2nd edition). Chicago: University of Chicago Press.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B, 50*, 157-224.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen, *Studies in social psychology in World War II, Volume 4: Measurement and prediction.* Princeton, NJ: Princeton university Press.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known item response function.* Paper presented at the Annual Meeting of the Psychometric Society, Nashville TN, June, 1985.

Lord, F.M. (1959). Statistical inference about true scores. *Psychometrika, 24,* 1-18.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Macready, G.B., & Dayton, C.M. (1989, March). *The application of latent class models in adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Marshall, S.P. (1985). *Using schema knowledge to solve story problems.* Paper presented at the Office of Naval Research Contractors' Conference, San Diego, CA, December, 1985.

Marshall, S.P. (1989). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Marshall, S.P. (in press). Assessing schema knowledge. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum.

Masters, G., & Mislevy, R.J. (in press). New views of student learning: Implications for educational measurement. In N. Frederiksen, R.J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum.

McDermott, L.C. (1984). Research on conceptual understanding in mechanics. *Physics Today, 37,* 24-32.

Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Mislevy, R.J. (in press-a). Randomization-based inference about latent variables from complex samples. *Psychometrika*.

Mislevy, R.J. (in press-b). Foundations of a new test theory. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Mislevy, R.J., & Sheehan, K.M. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27*, 255-272.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects follow different solution strategies. *Psychometrika, 55*, 195-215.

Novick, M.R. (1966). The axioms and principle results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Ohlsson, S., & Langley, P. (1985). Identifying solution paths in cognitive diagnosis. *Technical Report RI-TR-85-2*. Pittsburgh, PA: The Robotics Institute, Carnegie-Mellon University.

Paulsen, J.A. (1986). Latent class representation of systematic patterns in test responses. *Technical Report ONR-1*. Portland, OR: Psychology Department, Portland State University.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Kaufmann.

Rumelhart, D.A. (1980). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Erlbaum.

Siegler, R.S. (1981). Developmental sequences within and between concepts. *Monograph of the Society for Research in Child Development, 46*.

Snow, R.E., & Lohman, D.F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*, 347-376.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Tatsuoka, K.K. (1989). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tsutakawa, R.K., & Johnson, J. (1988). Bayesian ability estimation via the 3PL with partially known item parameters. *Mathematical Sciences Technical Report No. 147*. Columbia, MO: Department of Statistics, University of Missouri.

Wilson, M.R. (1989a). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education, 33*, 125-138.

Wilson, M.R. (1989b). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Yamamoto, K. (1987). *A model that combines IRT and latent class models*. Unpublished doctoral dissertation, University of Illinois.

Appendix

Equations for the Latent Class Model

The Model

Let $\eta = (\eta_0,...\eta_4)$ denote the stage of understanding of a child, with $\eta_k=1$ if he or she is in Stage k and 0 if not. Let $\pi = (\pi_0,...,\pi_4)$ denote the population proportions of children in these classes; that is, $\pi_k \equiv p(\eta_k=1)$. Let x_j represent a response to Task j, 1 if correct and 0 if not; j runs from 1 to 24. The conditional probabilities of correct response are $Prob(x_j=1|\eta_k=1)$, or P_{jk} for short. **P** denotes the matrix $((P_{jk}))$. A vector of item responses, $\mathbf{x} = (x_1,...,x_{24})$ is assumed to have the following probability *conditional on Stage membership*:

$$p(\mathbf{x}|\eta_k=1) = \prod_j P_{jk}^{x_j} (1-P_{jk})^{1-x_j} . \qquad (1)$$

Similar expressions are assumed to hold for subsets of responses as well, regardless of the order in which they are observed.

The *marginal* probability of a response vector is an average of terms like (1), weighted by the population probabilities of stage membership:

$$p(\mathbf{x}) = \sum_{k=0}^{4} p(\mathbf{x}|\eta_k=1)\ \pi_k. \qquad (2)$$

Let **X** denote the matrix of response vectors of a sample of N respondents. For a generic pattern $\mathbf{x}_\ell$, let n_ℓ be the number of respondents producing this pattern. The probability of **X** as a function of **P** and π has the form

$$P(\mathbf{X}|\mathbf{P},\pi) = C \prod_\ell p(\mathbf{x}_\ell)^{n_\ell} , \qquad (3)$$

where C does not depend on **P** or π. Once **X** has been observed, (3) can be interpreted as a likelihood function, and maxima may be found with respect to **P** and π.

Because N is only 120 in the balance beam example, a number of constraints were introduced so that stable estimates would be obtained. Many could be relaxed or removed

with larger samples. The results reported in Table 2 represent the best-fitting result among several models with similar numbers of constraints. The P_{jk}s that appear as .333 in Table 1 were fixed at that value. All four items of a given type were constrained to have the same P_{jk}s. For a given column, all P_{jk}s in cells that correspond to 1's in Table 1 were constrained to be equal to a single estimated value. Any cells in that column that correspond to 0's were constrained to its complement.

Adaptive Testing

The maximum likelihood estimates of **P** and π were treated as known true parameter values during simulated adaptive testing. The uncertainty in these values could be taken into account, but we have avoided the complication for this demonstration.

Before observing any responses from a given child, the expected value of his η is the population value π. The expected value of a response to a particular item j is obtained analogously to (2), simplified to a single, as yet unobserved, response:

$$p(x_j=1) = \sum_k p(x_j=1|\eta_k=1)\, p(\eta_k=1)$$

$$= \sum_k P_{jk}\, p(\eta_k=1)\,. \qquad (4)$$

Suppose that Item g is administered to a particular examinee, and the value of x_g, either 0 or 1, becomes known. How is this information propagated through the network? First, using Bayes theorem, we update probabilities for his η. For k=0,...,4,

$$p(\eta_k=1|x_g) = \frac{p(x_g|\eta_k=1)\, p(\eta_k=1)}{\sum_h p(x_g|\eta_h=1)\, p(\eta_h=1)}\,. \qquad (5)$$

This gives new probabilities that the examinee is in each of the possible stages. These are in turn reflected in new expectations for items not yet administered by replacing $p(\eta_k=1)$ in (4) with $p(\eta_k=1|x_g)$ to obtain

$$p(x_j=1|x_g) = \sum_k p(x_j=1|\eta_k=1)\, p(\eta_k=1|x_g)\,. \qquad (6)$$

This process can be repeated with additional items presented one at a time. Let $\mathbf{x}_s$ represent a partial response sequence; Item s+1 is next administered to form $\mathbf{x}_{s+1}$. Then

$$p(\eta_k=1|\mathbf{x}_{s+1}) = \frac{p(x_{s+1}=1|\eta_k=1)p(\eta_k=1|\mathbf{x}_s)}{\sum_h p(x_{s+1}=1|\eta_h=1)p(\eta_h=1|\mathbf{x}_s)} , \tag{7}$$

and, for items not yet presented,

$$p(x_j=1|\mathbf{x}_{s+1}) = \sum_k p(x_j=1|\eta_k=1)\ p(\eta_k=1|\mathbf{x}_{s+1}) . \tag{8}$$

Selecting which item to present next and deciding when to stop depends on probabilities for η. In this paper we have addressed only the case in which no decision-making cost structure is available, and we address only the goal of minimizing uncertainty about η. This can be accomplished by *minimum entropy* adaptive testing. Entropy is a measure of randomness. For the five-class balance beam problem, the maximal value of entropy occurs when probabilities of all five classes are equal, and the minimal value occurs when the probability of one particular stage is one. The general formula for entropy after having observed $\mathbf{x}_s$ is

$$E(\mathbf{x}_s) = -\sum_k p(\eta_k=1|\mathbf{x}_s) \log[p(\eta_k=1|\mathbf{x}_s)] . \tag{9}$$

After having observed $\mathbf{x}_s$, one can evaluate the expected entropy associated with the administration of any remaining item j as

$$E\left[\mathbf{x}_s \cap (x_j=0)\right] p(x_j=0|\mathbf{x}_s) + E\left[\mathbf{x}_s \cap (x_j=1)\right] p(x_j=1|\mathbf{x}_s) \tag{10}$$

The item that minimizes (10) is presented next.

It bears repeating that these formulae assume both that the model is correct and the conditional probabilities are known with certainty. Violations of these assumptions generally degrade knowledge about an examinee's state, making (5) and (8) in particular overly optimistic. Work remains to be done, in studying the robustness of the approach to violations of the assumptions, learning how to minimize violations in practice, and modifying the model or the conditional probabilities to mitigate inferential errors in the presence of violations.

TABLE 1

Theoretical Conditional Probabilities—
Expected Proportions of correct Response

Problem type	Stage 0	Stage I	Stage II	Stage III	Stage IV
E	.333	1.000	1.000	1.000	1.000
D	.333	1.000	1.000	1.000	1.000
S	.333	.000	1.000	1.000	1.000
CD	.333	1.000	1.000	.333	1.000
CS	.333	.000	.000	.333	1.000
CE	.333	.000	.000	.333	1.000

TABLE 2

Estimated Conditional Probabilities—
Expected Proportions of correct Response

Problem type	Stage 0	Stage I	Stage II	Stage III	Stage IV
E	.333*	.973	.883	.981	.943
D	.333*	.973	.883	.981	.943
S	.333*	.026	.883	.981	.943
CD	.333*	.973	.883	.333*	.943
CS	.333*	.026	.116	.333*	.943
CE	.333*	.026	.116	.333*	.943

* denotes fixed value

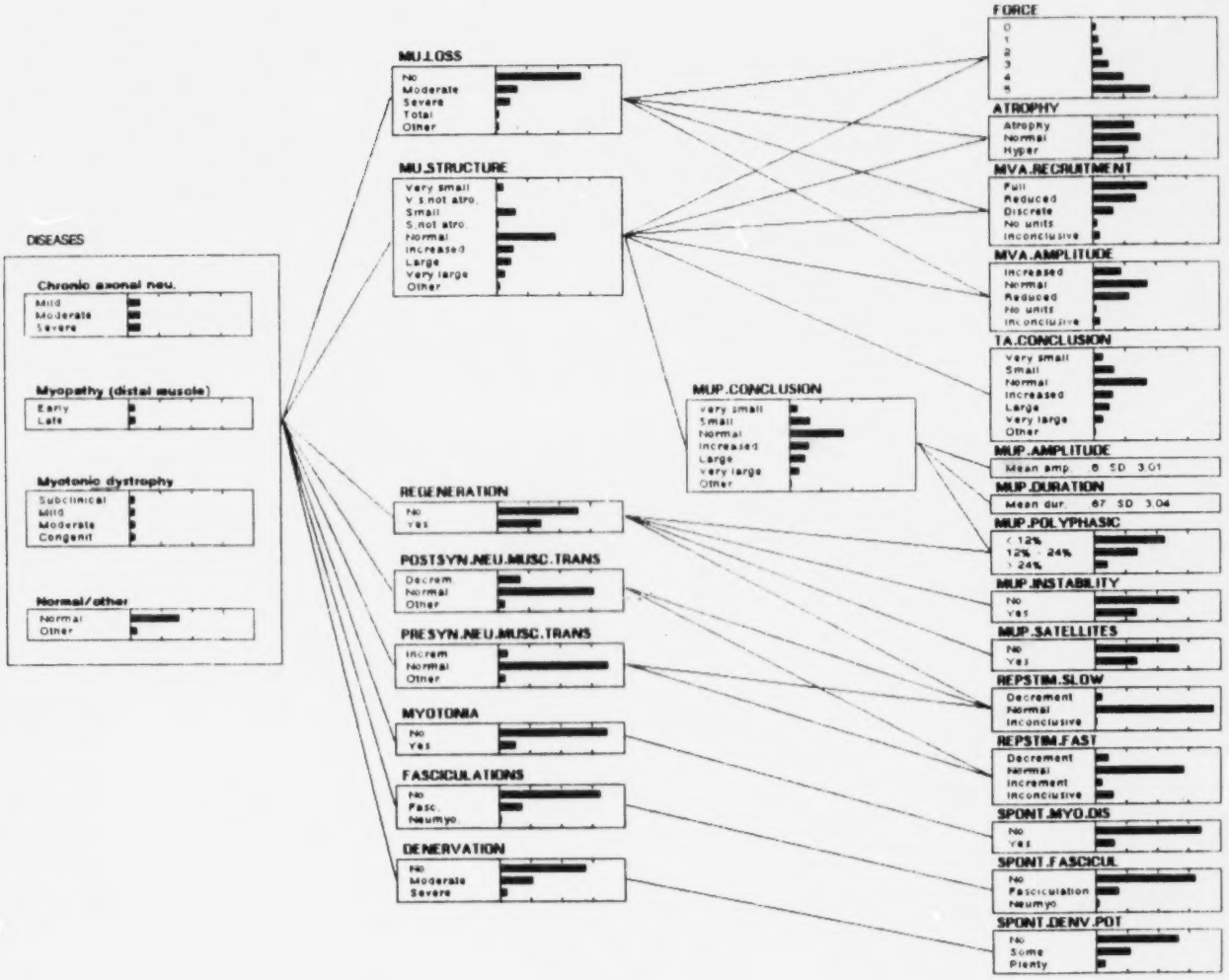


FIGURE 1

The MUNIN Network: Initial Status

(From Andreassen et al., 1987)

BEST COPY AVAILABLE

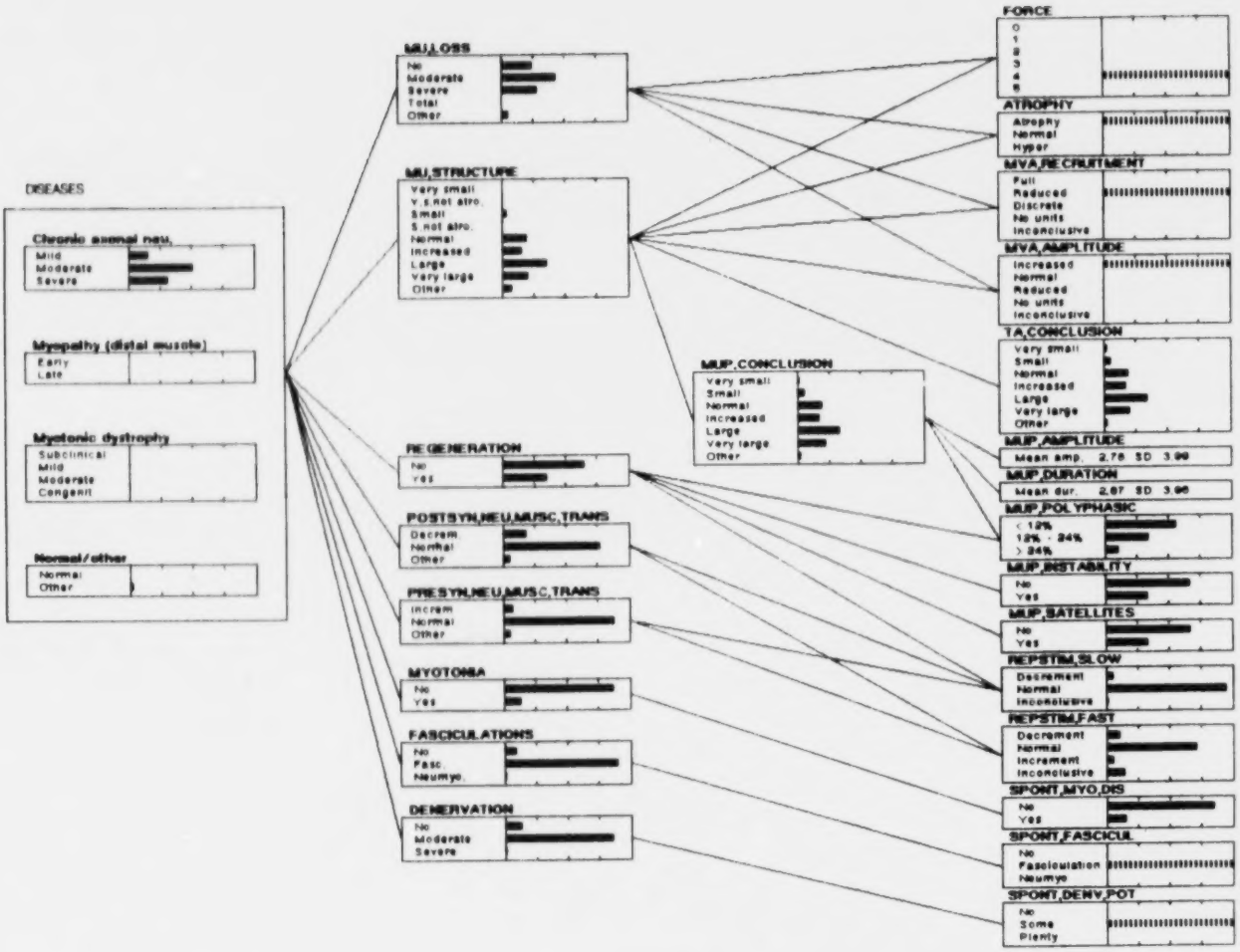
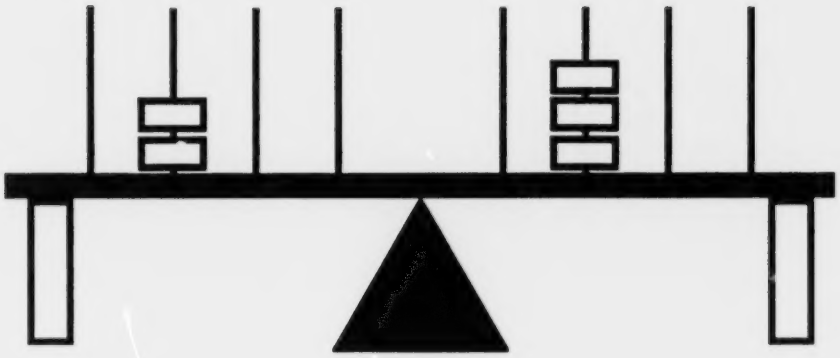


FIGURE 2

The MUNIN Network: After Selected Observations

(From Andreassen et al., 1987)

BEST COPY AVAILABLE



When the blocks are removed, will the beam tip left, tip right, or stay flat?

Figure 3

A Sample Balance-Beam Task

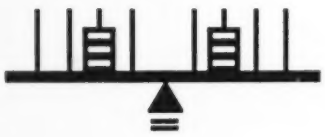
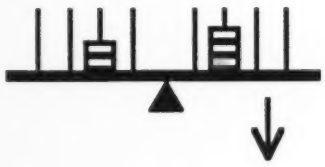
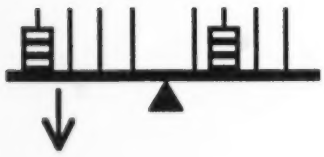
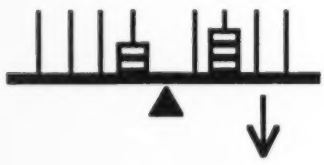
Item Type	Sample Item	Description
E		**Equal problems** (E), with matching weights and lengths on both sides.
D		**Dominant problems** (D), with unequal weights but equal lengths.
S		**Subordinate problems** (S), with unequal lengths but equal weights.
CD		**Conflict-dominant problems** (CD), in which one side has greater weight, the other has greater length, and the side with the heavier weight will go down.
CS		**Conflict-subordinate problems** (CS), in which one side has greater weight, the other has greater length, and the side with the greater length will go down.
CE		**Conflict-equal problems** (CE), in which one side has greater weight, the other has greater length, and the beam will balance.

Figure 4

Sample Balance Beam Items

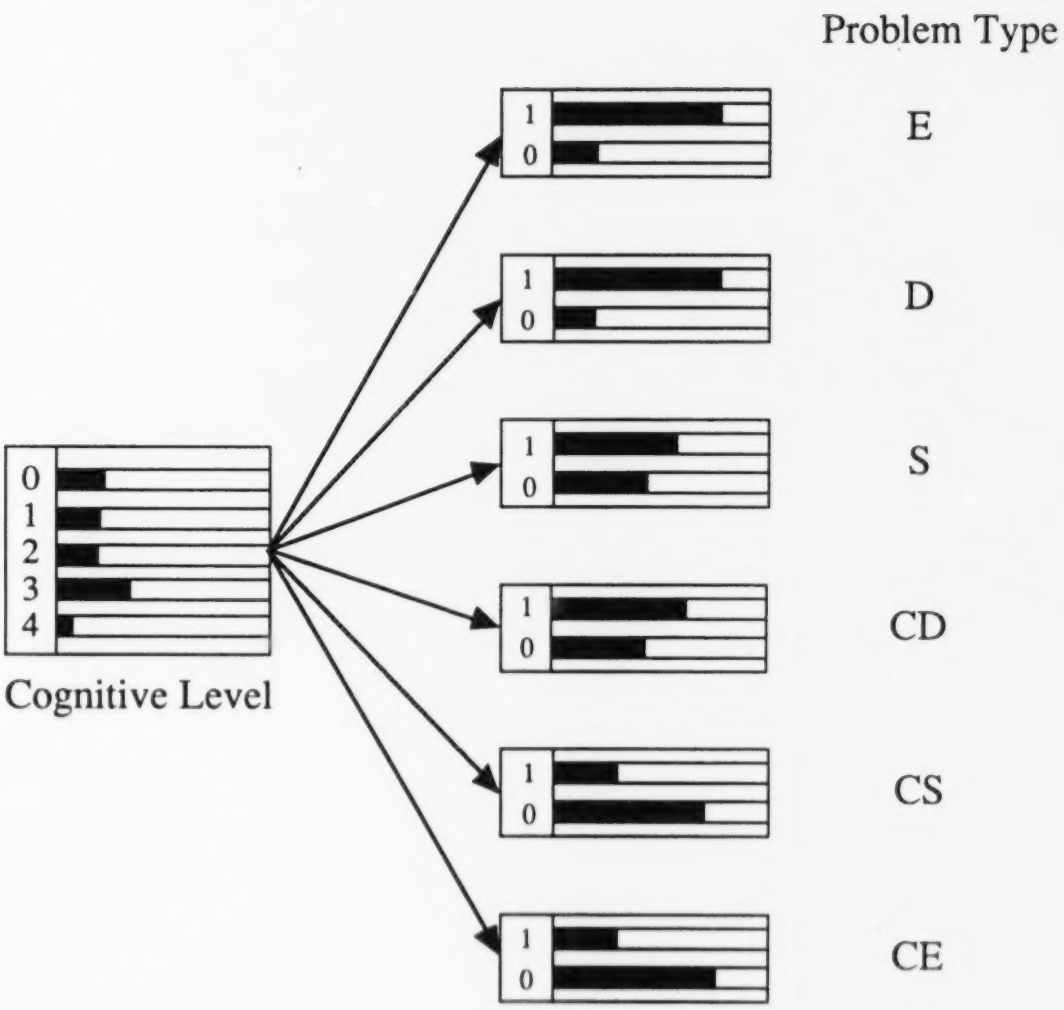


Figure 5
Initial State in an Inference Network
for the Balance Beam Example

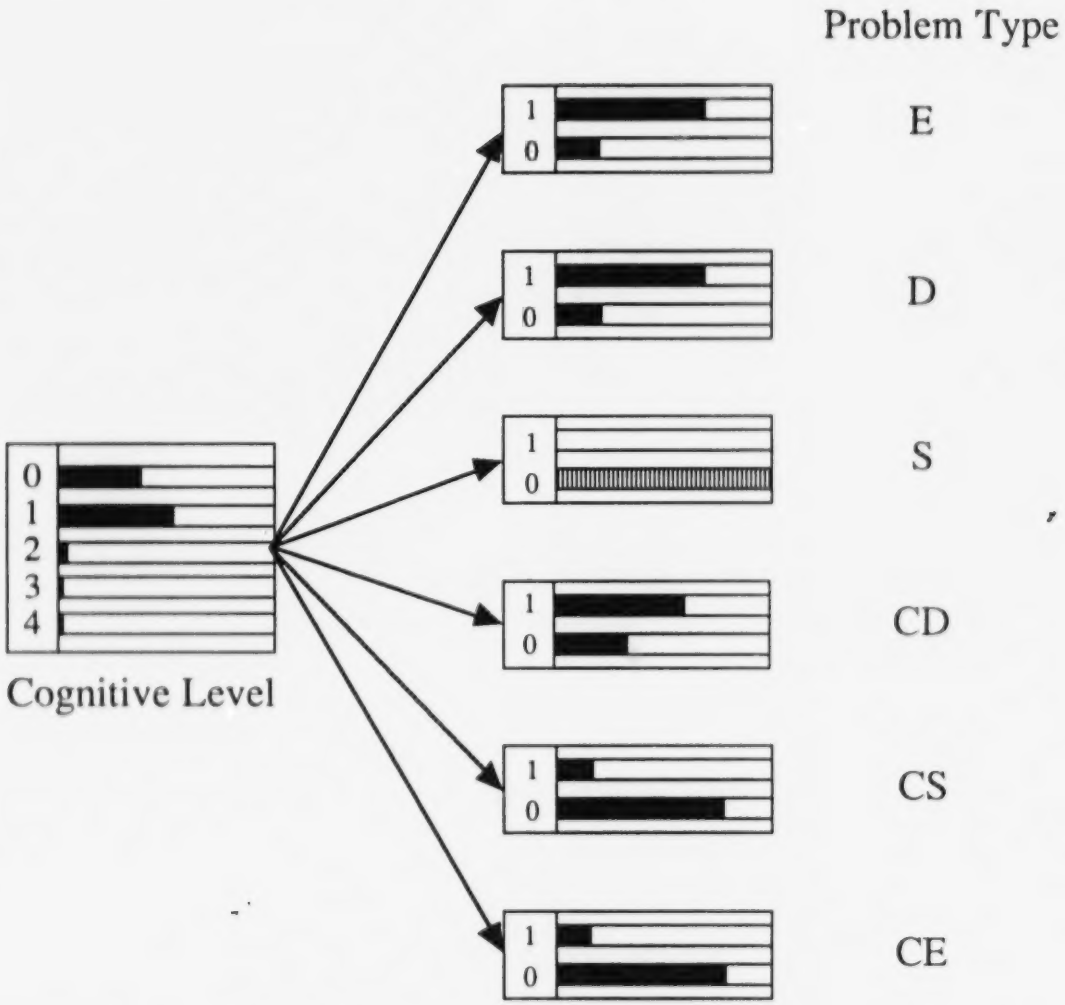


Figure 6

State of Knowledge about Cognitive Level

after an Incorrect Response to an S Item

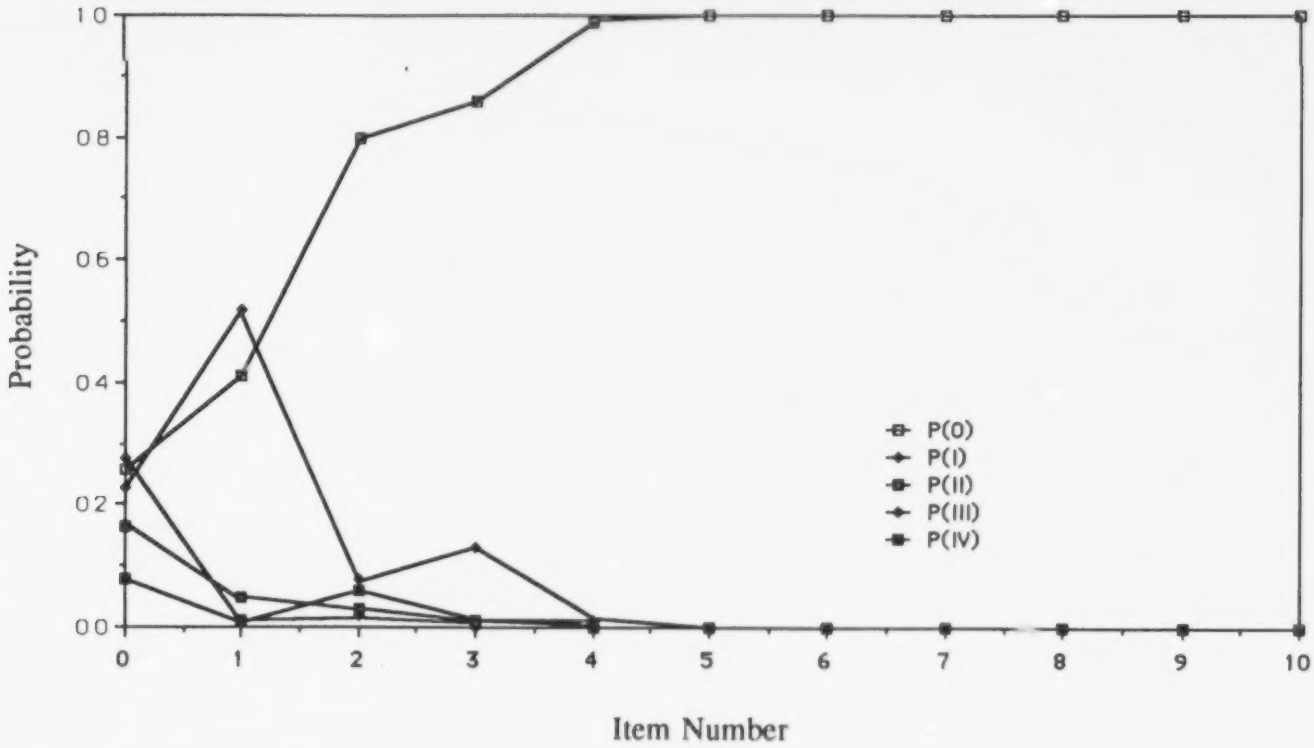


Figure 7

Posterior Probabilities of Cognitive Levels

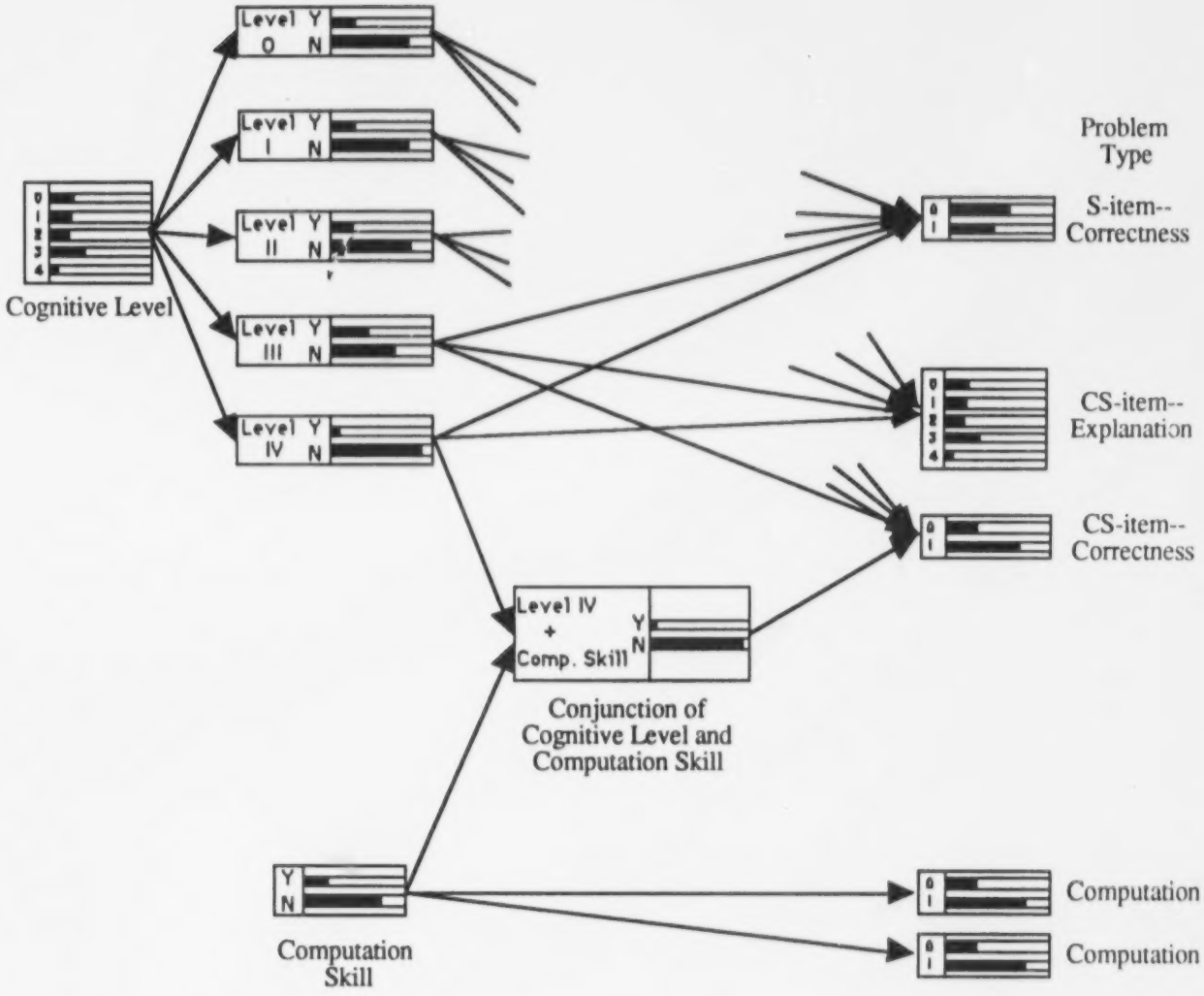


Figure 8

Representation of an Extended Balance-Beam Network

Dr. Terry Ackerman
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. James Algina
1403 Norman Hall
University of Florida
Gainesville, FL 32605

Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

Dr. Ronald Armstrong
Rutgers University
Graduate School of Management
Newark, NJ 07102

Dr. Eva L. Baker
UCLA Center for the Study
of Evaluation
145 Moore Hall
University of California
Los Angeles, CA 90024

Dr. Laura L. Barnes
College of Education
University of Toledo
2801 W. Bancroft Street
Toledo, OH 43606

Dr. William M. Bart
University of Minnesota
Dept. of Educ. Psychology
330 Burton Hall
178 Pillsbury Dr., S.E.
Minneapolis, MN 55455

Dr. Isaac Bejar
Mail Stop: 10-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Ramat Aviv 69978
ISRAEL

Dr. Arthur S. Blaiwes
Code N712
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Bruce Bloxom
Defense Manpower Data Center
99 Pacific St.
Suite 155A
Monterey, CA 93943-3231

Cdt. Arnold Bohrer
Sectie Psychologisch Onderzoek
Rekruterings-En Selectiecentrum
Kwartier Koningen Astrid
Bruijnstraat
1120 Brussels, BELGIUM

Dr. Robert Breaux
Code 281
Naval Training Systems Center
Orlando, FL 32826-3224

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. Gregory Candell
CTB/McGraw-Hill
2500 Garden Road
Monterey, CA 93940

Dr. John B. Carroll
409 Elliott Rd., North
Chapel Hill, NC 27514

Dr. John M. Carroll
IBM Watson Research Center
User Interface Institute
P.O. Box 704
Yorktown Heights, NY 10598

Dr. Robert M. Carroll
Chief of Naval Operations
OP-01B2
Washington, DC 20350

Dr. Raymond E. Christal
UES LAMP Science Advisor
AFHRL/MOEL
Brooks AFB, TX 78235

Mr. Hua Hua Chung
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director, Manpower Program
Center for Naval Analyses
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Director,
Manpower Support and
Readiness Program
Center for Naval Analysis
2000 North Beauregard Street
Alexandria, VA 22311

Dr. Stanley Collyer
Office of Naval Technology
Code 222
800 N. Quincy Street
Arlington, VA 22217-5000

Dr. Hans F. Crombag
Faculty of Law
University of Limburg
P.O. Box 616
Maastricht
The NETHERLANDS 6200 MD

Ms. Carolyn R. Crone
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Dr. Timothy Davey
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

Dr. C. M. Dayton
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Ralph J. DeAyala
Measurement, Statistics,
and Evaluation
Benjamin Bldg., Rm. 4112
University of Maryland
College Park, MD 20742

Dr. Lou DiBello
CERL
University of Illinois
103 South Mathews Avenue
Urbana, IL 61801

Dr. Dattprasad Divgi
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Mr. Hei-Ki Dong
Bell Communications Research
Room PYA-1K207
P.O. Box 1320
Piscataway, NJ 08855-1320

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Stephen Dunbar
224B Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. James A. Earles
Air Force Human Resources Lab
Brooks AFB, TX 78235

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Englehard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

Dr. Benjamin A. Fairbank
Operational Technologies Corp.
5825 Callaghan, Suite 225
San Antonio, TX 78228

Dr. P-A. Federico
Code 51
NPRDC
San Diego, CA 92152-6800

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-MRR
The Pentagon
Washington, DC 20310-0300

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

BEST COPY AVAILABLE

Dr. Alfred R. Fregly
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons
Illinois State Psychiatric Inst.
Rm 529W
1601 W. Taylor Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Drew Gitomer
Educational Testing Service
Princeton, NJ 08541

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Michael Habon
DORNIER GMBH
P.O. Box 1420
D-7990 Friedrichshafen 1
WEST GERMANY

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Grant Henning
Senior Research Scientist
Division of Measurement
Research and Services
Educational Testing Service
Princeton, NJ 08541

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 63
San Diego, CA 92152-6800

Dr. Thomas M. Hirsch
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 92010

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
3-104 Educ. N.
University of Alberta
Edmonton, Alberta
CANADA T6G 2G5

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign, IL 61820

Dr. Douglas H. Jones
1280 Woodfern Court
Toms River, NJ 08753

Dr. Brian Junker
Carnegie-Mellon University
Department of Statistics
Schenley Park
Pittsburgh, PA 15213

Dr. Milton S. Katz
European Science Coordination
Office
U.S. Army Research Institute
Box 65
FPO New York 09510-1500

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State
University
P.O. Box 522
Murfreesboro, TN 37132

Mr. Soon-Hoon Kim
Computer-based Education
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Mess. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. Richard J. Koubek
Department of Biomedical
& Human Factors
139 Engineering & Math Bldg.
Wright State University
Dayton, OH 45435

Dr. Leonard Kroeker
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Dr. Jerry Lehnus
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Mr. Rodney Lim
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Dr. Robert Lockman
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. Richard Luecht
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. George B. Macready
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. Clessen J. Martin
Office of Chief of Naval
Operations (OP 13 F)
Navy Annex, Room 2832
Washington, DC 20350

Dr. James R. McBride
The Psychological Corporation
1250 Sixth Avenue
San Diego, CA 92101

Dr. Clarence C. McCormick
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Mr. Christopher McCusker
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert McKinley
Educational Testing Service
Princeton, NJ 08541

Mr. Alan Mead
c/o Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Timothy Miller
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr. William Montague
NPRDC Code 13
San Diego, CA 92152-6800

Ms. Kathleen Moreno
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Headquarters Marine Corps
Code MPI-20
Washington, DC 20380

Dr. Ratna Nandakumar
Educational Studies
Willard Hall, Room 213E
University of Delaware
Newark, DE 19716

Dr. Harold F. O'Neil, Jr.
School of Education - WPH 801
Department of Educational
Psychology & Technology
University of Southern California
Los Angeles, CA 90089-0031

Dr. James B. Olsen
WICAT Systems
1875 South State Street
Orem, UT 84058

Dr. Judith Orasanu
Basic Research Office
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Dr. Peter J. Pashley
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

Dr. James Paulson
Department of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Malcolm Ree
AFHRL/MOA
Brooks AFB, TX 78235

Mr. Steve Reiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg.
Knoxville, TN 37916-0900

Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152-6800

Lowell Schoer
Psychological & Quantitative
Foundations
College of Education
University of Iowa
Iowa City, IA 52242

Dr. Mary Schratz
905 Orchid Way
Carlsbad, CA 92009

Dr. Dan Segall
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Robin Shealy
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr. Richard E. Snow
School of Education
Stanford University
Stanford, CA 94305

Dr. Richard C. Sorensen
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Judy Spray
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Dr. Peter Stoloff
Center for Naval Analyses
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

Mr. Brad Sympson
Navy Personnel R&D Center
Code-62
San Diego, CA 92152-6800

Dr. John Tangney
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Kikumi Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Maurice Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

Mr. Thomas J. Thomas
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Robert Tsutakawa
University of Missouri
Department of Statistics
222 Math. Sciences Bldg.
Columbia, MO 65211

Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

Dr. David Vale
Assessment Systems Corp.
2233 University Avenue
Suite 440
St. Paul, MN 55114

Dr. Frank L. Vicino
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Howard Wainer
Educational Testing Service
Princeton, NJ 08541

Dr. Michael T. Waller
University of Wisconsin-Milwaukee
Educational Psychology Department
Box 413
Milwaukee, WI 53201

Dr. Ming-Mei Wang
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Thomas A. Warm
FAA Academy AAC934D
P.O. Box 25082
Oklahoma City, OK 73125

BEST COPY AVAILABLE

Dr. Brian Waters
HumRRO
1100 S. Washington
Alexandria, VA 22314

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman
Box 146
Carmel, CA 93921

Major John Welsh
AFHRL/MOAN
Brooks AFB, TX 78223

Dr. Douglas Wetzel
Code 51
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Rand R. Wilcox
University of Southern
California
Department of Psychology
Los Angeles, CA 90089-1061

German Military Representative
ATTN: Wolfgang Wildgrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington, DC 20016

Dr. Bruce Williams
Department of Educational
Psychology
University of Illinois
Urbana, IL 61801

Dr. Hilda Wing
Federal Aviation Administration
800 Independence Ave, SW
Washington, DC 20591

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering
Cancer Center
1275 York Avenue
New York, NY 10021

Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
Code 51
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto
02-T
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Dr. Joseph L. Young
National Science Foundation
Room 320
1800 G Street, N.W.
Washington, DC 20550

Mr. Anthony R. Zara
National Council of State
Boards of Nursing, Inc.
625 North Michigan Avenue
Suite 1544
Chicago, IL 60611

BEST COPY AVAILABLE